OLAP Visualization Operator for Complex Data

Sabine Loudcher and Omar Boussaid

ERIC laboratory, University of Lyon (University Lyon 2)
5 avenue Pierre Mendes-France, 69676 Bron Cedex, France
Tel.: +33-4-78772320, Fax: +33-4-78772375
(omar.boussaid, sabine.loudcher)@univ-lyon2.fr

Abstract. Data warehouses and Online Analysis Processing (OLAP) have acknowledged and efficient solutions for helping in the decision-making process. Through OLAP operators, online analysis enables the decision-maker to navigate and view data represented in a multi-dimensional manner. But when the data or objects to be analyzed are complex, it is necessary to redefine and enhance the abilities of the OLAP. In this paper, we suggest combining OLAP and data mining in order to create a new visualization operator for complex data or objects. This operator uses the correspondence analysis method and we call it VOCoDa (Visualization Operator for Complex Data).
Keywords : OLAP, Data Mining, Complex Data, Visualization

1 Introduction

Data warehouses and Online Analysis Processing (OLAP) have recognized and effective solutions for helping in the decision-making process. Online analysis, thanks to operators, makes it possible to display data in a multi-dimensional manner. This technology is well-suited when data are simple and when the facts are analyzed with numeric measures and qualitative descriptors in dimensions. However, the advent of complex data has questioned this process of data warehousing and online analysis.

Complex data often contain a document, an image, a video, ..., and each of these elements can be described and observed by a set of low-level descriptors or by semantic descriptors. This set of elements can be seen not only as complex data but also as a complex object. A complex object is a heterogeneous set of data, which, when combined, form a semantic unit. For instance, a patient's medical record may be composed by heterogeneous elements (medical test results, X-rays, ultrasounds, medical past history, letter from the current doctor, ...) and is a semantic unit. It is a complex object.

As said above, warehousing and online analytical processes must be modified in the case of complex objects. In this paper, we focus on the visualization of complex objects. The problem of storing and modeling complex objects is discussed in other articles. The purpose of online analysis is to (1) aggregate many data to summarize the information they contain; (2) display the information according to different dimensions (3) navigate through data to explore them. OLAP operators are well-defined for classic data. But they are inadequate when data are complex. The use of other techniques, for example data mining, may be promising. Combining data mining methods with OLAP tools is an interesting solution for enhancing the ability of OLAP to analyze complex objects. We have already suggested extending OLAP capabilities with complex object exploration and clustering.

In this paper, we are concerned with the problem of the visualization of complex objects in an OLAP cube. By this means, we aim to define a new approach to extending OLAP capabilities to complex objects. With the same idea of combining data mining and online analysis, some works suggest using *Visual Data Mining* technology for visually and interactively exploring OLAP cubes. Maniatis *et al.* list possible representations for displaying a cube and offer the CPM model (*Cube Presentation Model*) as a model in an OLAP interface [3]. The CPM model borrows visualization tools from the field of the HMI (Human Machine Interface). Unfortunately, these works do not take complex objects into account . In a cube of complex objects, the facts are indeed complex objects, and the dimensions can include images, texts, descriptors, ... and OLAP measures are not necessarily numeric. Given these characteristics, standard visualization tools are not necessarily well-suited and should be adapted. To do this, we use the well-known principle of the factor analysis method in data mining. Factor analysis makes it possible to visualize complex objects while highlighting interesting aspects for analysis. This technique represents objects by projecting them on to factor axes. In a previous paper, we laid the foundations for this proposal [4]. In this paper, we complete and improve our first proposal by taking into account the measure to visualize complex objects, using indicators to make interpretation easier. We thus offer a comprehensive approach and a new OLAP operator entitled VOCoDa (*Visualization Operator for Complex Data*).

2 Running example

To illustrate our point of view, we complete the previously used case of researchers' publications. A publication can be seen as a complex object, or as a semantic entity. We plan to analyze publications according to their authors, national or international range, support such as a conference or a journal, etc. We aim to observe the diversity of the themes in which researchers publish and the proximity of authors when they are working on the same themes. Here, we observe publications as complex objects. To handle these semantic entities, we therefore need an adapted modeling and analysis tools.

In addition to standard descriptors such as year, type, authors, number of pages, etc., the user may also want to analyze the semantic content of the publication, i.e. the topics of the publication. The semantic content of the publication must be taken into account when modeling and carrying out an analysis. Let us suppose that the user wants to analyze publications according to the first author, support, year, content and topics of the paper.

The obtained multidimensional model is shown in figure 1.

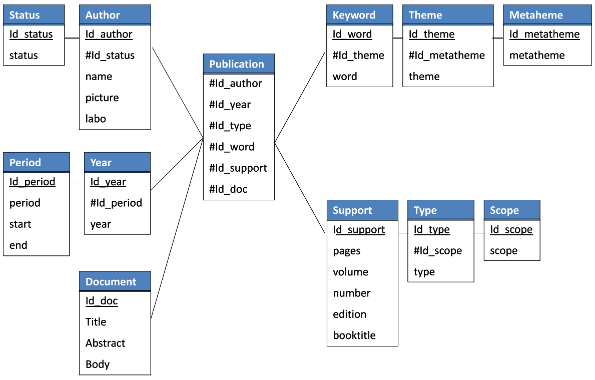

Fig. 1. Multidimensional modeling of publications

In this model, we believe that each dimension can be the fact and that objects are interchangeable in multi-dimensional modeling. There are therefore ”classic” dimensions with hierarchies, and semantic dimensions consisting of a hierarchy of concepts (keywords−− >themes−− >metathemes) and the document itself. Here, the fact is the publication and it is a combination of all dimensions without a measure. Generally, in case like this where there are no measures , the aggregation function COUNT can be used to count the facts. This solution is always possible in our case, but it is not sufficient because the analysis which follows is too poor. We seek other means to analyze publications in order to discover thematic proximity, authors who work together,... We consider a publication as a complex object and we are looking for a way to make a semantic analysis. We propose a visualization of complex objects which takes the semantic content of objects into account. This explains our decision to use a factor analysis method for the visualization of complex objects. This new visualization method fits completely with the online analysis of complex objects.

3 Positioning and principle

Generally, OLAP interfaces represent a cube as a table, or cross-table. In an attempt to exceed the limits of standard interfaces, more advanced tools offer visual alternatives to represent the information contained in a cube, and to interactively browse the cube (hierarchical visualizations, trees of decomposition, multi-scale views, interactive scatter plots) [8]. For a better visualization of information, Sureau *et al.* suggest rearranging the modalities of a level according to heuristics, based on distance between the elements in a dimension or according to a genetic algorithm [7]. With a statistical test, Ordonez and Chen searched within a cube (of low dimension) for neighboring cells with significantly different measures [6]. In the context of Web and OLAP applications, Aouiche *et al.* use a tag cloud to represent a cube where each keyword is a cell and where keyword size depends on the measured value of the fact (cell) [1].

Compared with the other approaches presented, we suggest a visualization operator (1) in the context of online analysis (2) that requires no assumptions about the data (3) that is suitable for complex objects (4) and that takes the semantic content of the data into account. Works on OLAP visualization do not deal with complex objects (even if some might be adapted to such data) and do not take the semantic content (only tag clouds seem to do this) into account.

To visualize complex objects, we propose an approach that uses factor analysis, a well-known method in data mining [2], [5]. A factor method makes it possible to visualize complex objects while highlighting interesting facts for analysis. When facts are complex objects, often there is no measure in the classical sense of multi-dimensional modeling. However, it is always possible to count the facts. In this case, the complex object cube with several dimensions with the COUNT function can be seen as a contingency table. Correspondence analysis (CA) can be used to display the facts. CA produces factor axes which can be used as new dimensions, called "factor dimensions". These new axes or dimensions constitute a new space in which it is possible to plot the facts i.e. complex objects. Using CA as the visualization operator is fully justified because this method has the same goal as OLAP navigation and exploration.

4 Process

We provide OLAP users with a process composed of several steps: (1) building the complex object cube, (2) constructing the contingency table, (3) completing the correspondence analysis, (4) mapping complex objects on the factorial axes.

Suppose that the user wants to study keywords in order to identify the major research fields in which researchers are working. In addition, the user would like to identify researchers working on the same keywords.

4.1 Notations

Let $\mathcal{C}$ be a cube with a non-empty set of d dimensions $\mathcal{D} = \{D^1, ..., D^i, ..., D^d\}$ and m measures $\mathcal{M} = \{M_1, ..., M_q, ..., M_m\}$. $\mathcal{H}^i$ is the set of hierarchical levels of dimension D^i. H^i_j is the j hierarchical level of dimension D^i. For example, the type of publication dimension D^1 has two levels: the level *Type* denoted H^1_1 and the level *Scope* denoted H^1_2.

$\mathcal{A}^{ij} = \{a^{ij}_1, ..., a^{ij}_t, ..., a^{ij}_l\}$ is the set of the l members or modalities a^{ij}_t of the hierarchical level H^i_j of the dimension D^i. The level *Scope* (H^1_2) has two members: *International*, denoted a^{12}_1 and *National*, denoted a^{12}_2.

4.2 Complex object cube

Depending on what the user wants to analyze, a cube is defined. This constructed cube is a sub-cube from the initial cube C. Let $\mathcal{D}'$ be a non-empty sub-set of $\mathcal{D}$ with p dimensions $\{D^1, ..., D^p\}$ ($\mathcal{D}' \subseteq \mathcal{D}$ and $p \leq d$). The p-tuple $(\Theta^1, ..., \Theta^p)$ is sub-cube if $\forall i \in \{1, ..., p\}$, $\Theta^i \neq \emptyset$ and if there is an unique $j \geq 1$ such that $\Theta^i \subseteq \mathcal{A}^{ij}$. A sub-cube, noted $\mathcal{C}'$, corresponds to a portion from the initial cube $\mathcal{C}$. Of the d existing dimensions, only p are chosen. For each chosen dimension $D^i \in \mathcal{D}'$, a hierarchical level H^i_j is selected and a non-empty sub-set Θ^i of members is taken from all the member set $\mathcal{A}^{ij}$ of the level.

For example, the user can choose to work in the context of the publications that were written between 2007 and 2009, by authors with the status of full professor. And in this context, the user can build, a cube of publications based on keywords, year of publication and the name of the first author. In our example, the sub-cube is given by $(\Theta^1, \Theta^2, \Theta^3, \Theta^4)$= ({*full professor*},{*2007, 2008, 2009*},{*Keyword 1, Keyword 2, ..., Keyword 4*},{*Author 1, Author 2, ..., Author 4*}). The measure M_q is the number of publications (*Count*).

4.3 Contingency table

Classically, correspondence analysis takes as input a contingency table. Our idea is to use traditional OLAP operators to build this contingency table.

In the sub-cube $\mathcal{C}'$, the user chooses two levels (one level for two different dimensions), on which he wants to visualize complex objects. Let Θ^i (respectively $\Theta^{i'}$) be the set of l (respectively l') members chosen for the level of the dimension i (respectively i'). The contingency table $\mathcal{T}$ has l rows and l' columns the titles of which are given by $\{a^{ij}_1, ..., a^{ij}_t, ..., a^{ij}_l\}$ and $\{a^{i'j'}_1, ..., a^{i'j'}_{t'}, ..., a^{i'j'}_{l'}\}$. At each intersection of row t and column t', are counted the facts having the members a^{ij}_t and $a^{i'j'}_{t'}$.

In our example, the contingency table crosses keywords with authors in the sub-cube. This consists in counting facts covering 3 years by doing a roll-up of the dimension *year*. This gives us a cross table with keywords in rows and authors in columns. At the intersection of a row and a column, we have the number of publications written by an author for a given keyword. This table is ready to be processed by a CA. If the measure used is other than a simple count, and if it is a numerical measure, additive and with only positive values, then it is possible to use it to weigh the facts in the contingency table. The user is given the choice of using this measure as weighting or not.

4.4 Correspondence analysis

Processing a CA consists in projecting data on to synthetic axes so that much information is expressed by a minimum number of axes. The goal is to reduce the size of the representation space, that is to say, to reduce the number of rows and columns. The CA makes possible simultaneous visualization of the projections of rows and columns in the same plane. The proximities between rows and columns can be interpreted.

In practice, the method starts by calculating the eigen values from which are deduced eigen vectors that define the factor axes. As the first two axes contain the most information, they define the first factor plane. Once row points and column points have been projected on to axes, auxiliary statistics are reported to help evaluate the quality of the axes and their interpretation. For each point, the most important statistics are the weight, the relative contribution of the point to the axis' inertia and the quality of the representation on the axis (given by the $cosine^2$). To give an interpretation of an axis and analyze proximity between points on an axis, only points which contribute strongly to the inertia of the axis (whose contribution is three times the average contribution) and which are well represented by the axis (whose $cosine^2$ is higher than 0.5) are taken into account.

4.5 Visualization

The first two factor axes are retained as new factor dimensions, because the coordinates of the projected objects can be seen as members of dimensions. The graph in figure 2 is obtained. It allows representing publications according to their semantic content described by authors and keywords. It is possible to interpret the factor dimensions. Once the graph has been constructed, an interactive tool gives, for each point, i.e. keyword or author, its statistic indicators (relative contribution and $cosine^2$). Keywords and authors that have high indicators are represented in a different color. Thus, the user sees the most relevant points for analysis. Factor analysis provides automatic help in understanding and to analyzing information. For example, the user can easily identify the most characteristic keywords, authors who work together or who do not work together and finally groups of authors working on certain keywords. In addition, if the user so requests, a photograph of the authors can replace their name. In an OLAP framework, it is efficient to use the most significant descriptors of dimensions in order to enhance the readability of the results obtained.

Furthermore, according to the OLAP principle, it is also possible on each point to perform a *drill-down* to see related publications (represented by their title). The user has another possibility of projecting a hierarchical level of another dimension into the graph. The members of this new level will be projected as points in factor space but they have not been involved in the construction of the axes. To maintain statistical consistency, only hierarchical levels whose dimensions are not in the sub-cube can be used as additional elements. A level of a dimension already used would be dependent on another level. In our example, the user could use as an additional element type of publication (journal, conference, technical report ...).

We have developed a software platform implemented as a Web Open Source application in *PHP5* and with a *MySQL* database. It uses the *R* software and its *FactoMiner* package. The graphic interface is managed by an *ExtJS* framework with an *Ajax* support.

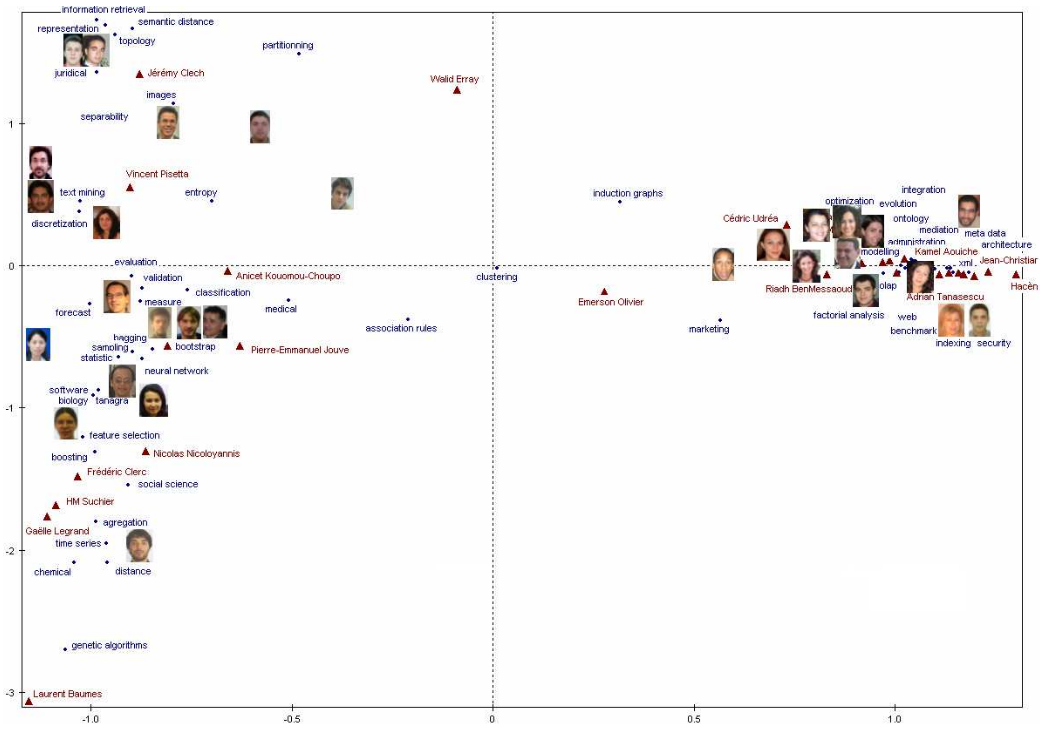

Fig. 2. Visualization of publications

5 Conclusion

In this paper, we have developed an approach to online analysis for complex objects. Our approach has demonstrated the feasibility of using correspondence analysis to make it possible to visualize complex objects online taking their semantic content into account. Furthermore, it naturally takes its place in the online analysis. The publications case study illustrates our approach. In the proposed multi-dimensional model, publications are described by keywords. Rather than asking authors to assign keywords themselves manually to their publication or rather than using an ontology, we think that it would be more relevant to automatically extract the keywords from the title, summary, or text (body) of the publication. Indeed, if the keywords were automatically extracted, they would capture some of the semantics contained in the document. Using information retrieval (IR) principles, keywords could be extracted automatically. Furthermore, as publications contain documents and documents contain text, our idea is to use certain information retrieval (IR) techniques in order to model publications. The use of IR techniques can allow us to extract semantics from the text and this semantic information may be very helpful for modeling publications in a multi-dimensional manner. In addition to combining OLAP and data mining, the coupling of OLAP and IR should further enhance online analysis.

References

1. K. Aouiche, D. Lemire and R. Godin. Web 2.0 OLAP: From Data Cubes to Tag Clouds. Proceedings of the 4^{th} International Conference on Web Information Systems and Technologies (WEBIST 08). 2008, 5–12.
2. J.P. Benzecri. Correspondence Analysis Handbook. Marcel Dekker, hardcover edition, 1992.
3. A. S. Maniatis, P. Vassiliadis, S. Skiadopoulos, Y. Vassiliou. Advanced visualization for OLAP. Proceedings of the 6^{th} ACM International Workshop on Data Warehousing and OLAP (DOLAP'2003). 2003,9–16.
4. L. Mabit, S. Loudcher, O. Boussaid. Analyse en ligne d'objets complexes avec l'analyse factorielle. 10^{me} Confrence d'Extraction et Gestion des Connaissances (EGC 2010). 2010, 381–386.
5. M. Greenacre. Correspondence Analysis in Practice. Chapman Hall CRC, Second Edition. 2007.
6. C. Ordonez, Z. Chen. Exploration and Visualization of OLAP Cubes with Statistical Tests. Proceedings of the 15^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Workshop on Visual Analytics and Knowledge Discovery. 2009,46–55.
7. F. Sureau, F. Bouali, G. Venturini. Optimisation heuristique et gntique de visualisations 2D et 3D dans OLAP : premiers rsultats. 5mes Journes francophones sur les Entrepts de Donnes et l'Analyse en ligne (EDA 09). 2009, 62–75.
8. S. Vinnik, F. Mansmann. From analysis to interactive exploration: Building visual hierarchies from OLAP cubes. Proceedings of the 14^{th} International Conference on Extending Database Technology (EDBT'2006). 2006, 496–514.